PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 7 : G10L

A2

(11) International Publication Number: **WO 00/54252**

(43) International Publication Date: 14 September 2000 (14.09.00)

(21) International Application Number: PCT/EP00/01145

(22) International Filing Date: 10 February 2000 (10.02.00)

(30) Priority Data:
199 10 234.1 9 March 1999 (09.03.99) DE

(71) Applicant: KONINKLIJKE PHILIPS ELECTRONICS N.V. [NL/NL]; Groenewoudseweg 1, NL–5621 BA Eindhoven (NL).

(71) Applicant *(for DE only)*: PHILIPS CORPORATE INTELLECTUAL PROPERTY GMBH [DE/DE]; Habsburgerallee 11, D–52066 Aachen (DE).

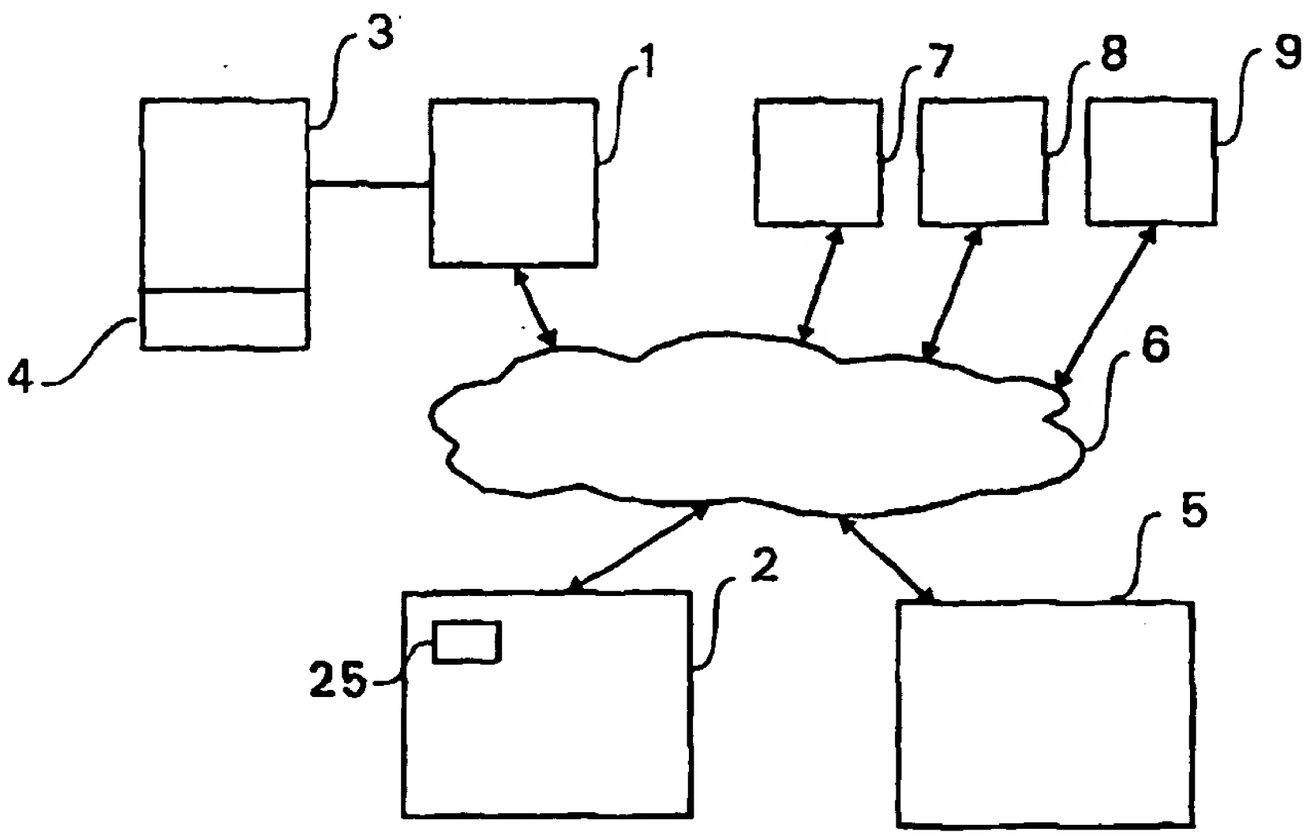
(72) Inventors: ULLRICH, Meinhard; Prof. Holstlaan 6, NL–5656 AA Eindhoven (NL). THELEN, Eric; Prof. Holstlaan 6, NL–5656 AA Eindhoven (NL). BESLING, Stefan; Prof Holstlaan 6, NL–5656 AA Eindhoven (NL).

(74) Agent: VOLMER, Georg; Prof. Holstlaan 6, NL–5656 AA Eindhoven (NL).

(81) Designated States: AE, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE, DK, DM, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: METHOD WITH A PLURALITY OF SPEECH RECOGNIZERS



(57) Abstract

The invention relates to a method in which an information unit (3) that makes a speech input possible is stored on a server (1) and can be retrieved by a client (2) and the client (2) can be coupled through the communications network (6) to a plurality of speech recognizers (7–9) and a speech input given by a user is applied to at least one speech recognizer (7–9) for generating at least one recognition result (11–13) and the recognition result (11–13) is interpreted in a plurality of independent processes and a plurality of interpretation results (22–24) are generated which are sent to a user. The user then receives in a brief period of time a plurality of qualified information items for which otherwise he would several times have had to make an inquiry in databases by means of a speech input.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania	**ES**	Spain	**LS**	Lesotho	**SI**	Slovenia
AM	Armenia	**FI**	Finland	**LT**	Lithuania	**SK**	Slovakia
AT	Austria	**FR**	France	**LU**	Luxembourg	**SN**	Senegal
AU	Australia	**GA**	Gabon	**LV**	Latvia	**SZ**	Swaziland
AZ	Azerbaijan	**GB**	United Kingdom	**MC**	Monaco	**TD**	Chad
BA	Bosnia and Herzegovina	**GE**	Georgia	**MD**	Republic of Moldova	**TG**	Togo
BB	Barbados	**GH**	Ghana	**MG**	Madagascar	**TJ**	Tajikistan
BE	Belgium	**GN**	Guinea	**MK**	The former Yugoslav Republic of Macedonia	**TM**	Turkmenistan
BF	Burkina Faso	**GR**	Greece			**TR**	Turkey
BG	Bulgaria	**HU**	Hungary	**ML**	Mali	**TT**	Trinidad and Tobago
BJ	Benin	**IE**	Ireland	**MN**	Mongolia	**UA**	Ukraine
BR	Brazil	**IL**	Israel	**MR**	Mauritania	**UG**	Uganda
BY	Belarus	**IS**	Iceland	**MW**	Malawi	**US**	United States of America
CA	Canada	**IT**	Italy	**MX**	Mexico	**UZ**	Uzbekistan
CF	Central African Republic	**JP**	Japan	**NE**	Niger	**VN**	Viet Nam
CG	Congo	**KE**	Kenya	**NL**	Netherlands	**YU**	Yugoslavia
CH	Switzerland	**KG**	Kyrgyzstan	**NO**	Norway	**ZW**	Zimbabwe
CI	Côte d'Ivoire	**KP**	Democratic People's Republic of Korea	**NZ**	New Zealand		
CM	Cameroon			**PL**	Poland		
CN	China	**KR**	Republic of Korea	**PT**	Portugal		
CU	Cuba	**KZ**	Kazakstan	**RO**	Romania		
CZ	Czech Republic	**LC**	Saint Lucia	**RU**	Russian Federation		
DE	Germany	**LI**	Liechtenstein	**SD**	Sudan		
DK	Denmark	**LK**	Sri Lanka	**SE**	Sweden		
EE	Estonia	**LR**	Liberia	**SG**	Singapore		

Method with a plurality of speech recognizers.

The invention relates to a method in which an information unit that makes a speech input possible is stored on a server and can be retrieved by a client.

The possibility of carrying out the communication with a computer by speech input instead of keyboard or mouse unburdens the user in his work with computers and often increases the speed of input. Speech recognition can be used in many fields in which nowadays data input is effected by keyboard.

EP 0 872 827 describes a system and a method of speech recognition. A client on which compressed software for speech recognition is executed is connected to a speech recognition server through a network. The client sends a speech recognition grammar and the data of the speech input to the speech recognition server. The speech recognition server executes the speech recognition and returns the recognition result to the client.

When a user is interested in information, he looks for this information at the location known to him. The fact that there are more than one service providers for a certain area is often unknown to the user. Different service providers respond differently to the user's respective inquiries. Mostly, however, the user does not know where a further information source exists. Even if he knew, he would have to make a new inquiry. This is time-consuming.

Therefore, it is an object of the invention to give the user as much qualified information as possible in a brief period of time.

This object is achieved in that the client can be coupled through a communications network to a plurality of speech recognizers and a user's speech input is applied to at least one speech recognizer for the generation of a recognition result and the recognition result is interpreted in a plurality of independent processes and a plurality of interpretation results are generated which are supplied to the user.

A service provider stores an information unit on a server, which information unit makes a speech input possible. A client downloads an information unit from this server, which information unit makes a speech input possible. A server is a computer in a communications network, for example, the Internet, on which information of providers is stored and can be retrieved by clients. A client is a computer which is connected to a server for retrieving information from the Internet and downloads the information unit stored on the

server to represent the information unit by means of software. This information unit is delivered by the client so that the user can perceive the contents of this information unit. The user is requested either by the information unit to enter speech, or, since this information unit has often been invoked, is informed about the possibility of entering speech. After the user has given a speech input, this speech input is applied to one or more speech recognizers. The individual speech recognizers execute a speech recognition and each generate a recognition result. These recognition results are each subjected to an interpretation. The recognition results are used to come to interpretation results in independent processes. For an interpretation of a recognition result, this recognition result is analyzed. Therefore, the recognition result is subdivided into its component parts and for example keywords are looked for. Parts of the recognition result that are uninteresting for a later information inquiry are omitted. The analysis can then be made from the speech recognizer or from a database. For analyzing the recognition result it is therefore necessary to have information about the contents of the speech input. Possible contents of the speech input are determined by the contents of the information unit. By means of this analysis, an inquiry is made for a database. This inquiry is then sent to the individual databases which thereafter produce a plurality of independently generated interpretation results. An important aspect which has a decisive influence on the quality of the response to the speech input made by the user is the database which is used for finding an answer to an inquiry. The number of independent databases is ever rising. Furthermore, there are extensive databases of businesses which may also assist in finding an answer. These separate databases are integrated in that the recognition results are assigned to the databases for multiple interpretation when answers are to be found.

The speech recognition for generating recognition results can be used with different cost levels. Speech recognizers are distinguished not only by their size and specialization of the vocabulary, but also by the algorithms with which they perform the speech recognition. A good database inquiry requires a good recognition of this inquiry made by the user via his speech input.

The interpretation results from the speech recognizer or the database are either automatically sent back to the client, or the server renders them available, so that the user can retrieve the individual interpretation results as required. In either case the interpretation results are delivered by the client in a form that can be perceived by the user.

Due to the combination of the information unit and one or more speech recognizers, the user is provided with a multiple answer to his inquiry made by speech input.

As a result, he receives information for which, without this method, he would have to start more than one inquiries with considerable time delay.

Apart from different recognition results during the speech recognition, different interpretation results are generated as a result of the independent interpretation of the individual recognition results based on different databases, which interpretation results each give a response to the speech input coming from the user. With a single interpretation of the speech input, either only a limited number of the most probable answers to the inquiry would be sent back to the client, or the user would receive responses which are much beside the inquiry as regards their contents. As a result of the multiple interpretation of one or more recognition results, the user is informed of at least double the amount of information in the same time.

When the speech input is assigned to only one speech recognizer, the recognition result is fed to a plurality of interpretation processes which all produce an interpretation result which is sent back to the client or retrieved by him and thus provides a multiple response to the user's inquiry.

In a further embodiment of the invention it has proved to be advantageous to preprocess the speech input on the side of the client. For this purpose, additional software is started on the client when the information unit is loaded, which additional software carries out an extraction of the features of the speech input. This additional software digitizes, quantizes and subjects the speech input available as an electric signal to respective analyses, which produce components to which feature vectors are assigned. These feature vectors are then transmitted to the coupled speech recognizer. The speech recognizer executes the compute-intensive recognition. As a result of the extraction of the features executed on the client, the speech input is compressed and coded, so that the number of data to be transmitted is reduced. Furthermore, the time necessary for the feature extraction is reduced on the side of the client, so that the speech recognizer only executes the recognition of the feature vectors applied to it. With speech recognizers that are used frequently, this reduction may be advantageous. When the speech input is assigned to a plurality of speech recognizers, there is the advantage that the preprocessing needs to be carried out only once. Without the feature extraction on the side of the client, each selected speech recognizer would execute such an extraction.

As a further embodiment of the invention, there is proposed that the client downloads the information unit in the form of an HTML page (Hyper Text Markup Language) from the server. This HTML page is shown by means of a Web browser on the client. The client sets up a connection by means of a link to the server, on which link the HTML page, in

which the user is interested in, is stored. The HTML page can contain graphic symbols, audio and/or video data in addition to text to be represented. The HTML page requests the user via an indication to make a speech input. After the user has made this speech input, this speech input is transferred from the client to one or more speech recognizers. A speech recognition is then executed there. The quality of the recognition result then decisively depends on how specialized the speech recognizers are. Speech recognizers work with a certain finite vocabulary, which is mostly limited to special fields of application. Therefore, it is important for a usable recognition result that the speech recognizers to which the speech input is transferred are accordingly specialized. The recognition result or a plurality of recognition results, as the case may be, is/are subjected to an interpretation process. For this purpose, for example the recognized speech input is analyzed for a database and on the basis of this analysis an inquiry is made to the data file of this database. The resulting interpretation result is automatically sent back to the client or retrieved by the client and represented there by a Web browser. The user can now make a choice from the plurality of interpretation results. This operation can be compared with looking up in a plurality of lexicons, with the advantage of saving time.

In a further embodiment of the invention there is provided to represent a plurality of objects, for example, commercials of firms on an HTML page, which each make a speech input possible. To each object is assigned a speech recognizer connected through the communications network, to which recognizer the speech input coming from the user is sent. The speech recognizers execute the speech recognition and convey the individual recognition results to independent interpretation processes. The interpretation results sent back to the client or retrieved by him are offered to the user in the form of a graphical representation or as an audio signal.

If the objects, which may be realized, for example, as advertising banners are offered by companies working in the same line of business, a user is presented with a plurality of offers from competing firms as a result of his speech input and its multiple parallel processing.

With advertising banners of non-competing firms, which are shown on an HTML page, a user's speech input relating to a specific advertising banner is conveyed to the speech recognizers assigned to an object in that the advertising banner is clicked on with the mouse or in that the user's point of vision is followed, or in that priorities are given to the plurality of speech input options of the individual objects. It is then advantageous to either store the speech input or the preprocessed speech input in a memory on the client, or to send

the recognition result back to the client, so that for the purpose of another interpretation process the user can employ this intermediate result which is available anyway. The stored speech input or recognition result is then conveyed to another speech recognizer if a speech input has been stored, or to another database if a recognition result has been stored, so as to be capable of making further interpretation results with further interpretations.

In a further embodiment a choice is made from a plurality of objects represented by the Web browser which are enabled by a speech input. From the total number of objects shown, the user chooses several objects, for example, by clicking the mouse. A speech input is then sent only to the speech recognizers of these chosen objects.

In a further embodiment of the invention, a server assigns additional information in the form of an HTML tag to each object to combine the object with a speech recognizer. As a result, while the HTML page is being downloaded, the object is informed of which speech recognizer on the Internet the speech input is to be sent to be processed.

Furthermore, with this additional information it is also possible to assign the databases on which the interpretation of the recognition results is to be effected. As a result, the provider of the HTML page determines to which database the recognition result or the inquiry is to be sent.

A further advantageous embodiment of the invention is provided by the possibility of leaving the decision to which databases the recognition result is sent up to the speech recognizer. This achieves a shift of the decision on which database the user's inquiry is to be processed. When the HTML page provider who assigns the speech recognizer to the respective object is not up to date as regards the databases, but the operator of the speech recognizers is and he is the one who assigns the databases, the quality of the response to the request is enhanced as a result thereof.

With an HTML page which informs about new publications of books and to which are switched a plurality of advertising banners of different publishers, the HTML page provider who is independent of publishers can send a recognition result from a user's inquiry about new publications in a respective field to all the databases available to him. As a result, the user rapidly receives extensive information about new publications of books of a respective field.

Furthermore, the object is also achieved by a server on which an information unit is stored which can be retrieved by a client, while there is provided that

- the client can be coupled to one or more speech recognizers for generating a plurality of interpretation results sent to a user, and

- a speech input is applied to at least one speech recognizer for generating recognition results and the recognition results are interpreted in a plurality of independent processes, and
- for determining a combination of an object that makes a speech input possible with a speech recognizer for generating a recognition result, additional information is assigned to the object.

These and other aspects of the invention are apparent from and will be elucidated with reference to the embodiments described hereinafter.

In the drawings:

Fig. 1: shows a block diagram of an arrangement for implementing the method according to the invention,

Fig. 2: shows a block diagram of the method according to the invention with a speech recognizer,

Fig. 3: shows a block diagram of the method according to the invention with parallel speech recognizers and

Fig. 4: shows a block diagram of the method according to the invention with parallel speech recognizers with an integrated database.

Fig. 1 shows by way of example an arrangement for implementing the method according to the invention. An information unit 3 is stored on a server 1. The server 1 can be coupled to a client 2 through a communications network 6. Through this communications network 6, called Internet 6 hereinafter, speech recognizers 7-9 can be coupled to the client 2. Also through the Internet 6, databases 5 can be coupled to the client 2, to the speech recognizers 7 and 9 and to the server 1.

A provider stores the information unit 3 on the server 1 to allow a user to access information, for example, via this provider. The information unit 3 contains not only contents to be represented and formatting instructions, but also additional information 4. The user downloads an information unit 3 which is of interest to him, in the following to be referenced HTML page 3, from the server 1. For this purpose, a connection based on the TCP/IP protocol is set up to the server 1. Software is executed on the client 2, which software may be realized, for example, by a Web browser and by which the HTML page 3 is shown to the user. The client 2 includes a memory 25 in which a speech input uttered by the user or a recognition result sent back by a speech recognizer 7-9 is stored.

Fig. 2 shows the information unit 3 which offers the user interactivity in the form of a speech-input option. The objects 19, 20 and 21 are advertising banners, which show

the user, for example, advertisements of car firms. Furthermore, they show the user that this HTML page 3 offers a speech input option in that the user, for example, by flashing text - for example, "tell us which car you are interested in" -, utters a speech input. In this example of embodiment all three advertising banners 19, 20, 21 expect to receive a similar speech input. Therefore, the speech input is conveyed to only one speech recognizer 7 via the Internet 6. For example, for finding a car, the user can pronounce concepts or word groups of interest to him, which are fed to the client by means of an input device 10 and are conveyed to the speech recognizer 7. By means of additional software (not shown), an extraction of the features of a speech input can be made on the client 2, so that the speech recognizer 7 is only supplied with the speech-input features arranged in feature vectors in compressed form. The speech recognizer 7 carries out the speech recognition and generates a recognition result 11. This recognition result 11 is analyzed and sent as an inquiry from the speech recognizer 7 to the databases 14, 15 and 16. The inquiries, which are in this case sent to the databases 14, 15 and 16, are the same.

The databases may also be located on the same server as the speech recognizer 7. However, it is also conceivable for the inquiries to be sent to databases which are located on different servers. It is then to be observed that the speech recognizer 7 belongs to the provider of the HTML page 3 or is hired by him. Since the provider knows that inquiries are made after cars on this HTML page 3, the client is connected to a specialized speech recognizer for recognizing the speech input. The database 14 contains data from a file of the car firm of advertising banner 19. Database 15 contains data of the car firm with advertising banner 20 and the database 16 of the car firm with advertising banner 21. The databases 14, 15 and 16 are then searched for information that is in line with the inquiry. This operation is also referenced interpretation. The databases 14, 15 and 16 each produce an interpretation result 22, 23 and 24 which is shown on the client 2 after being transmitted via the Internet 6. Together with the interpretation result 22 is presented to the user an offer from the car firm having advertising banner 19, with the interpretation result 23 an offer from the car firm having advertising banner 20 and with interpretation result 24 an offer from the car firm having advertising banner 21.

In this manner, information from three different databases 14-16 is rendered available to the user. He now receives, for example, an offer of a car from the file of the car firm having advertising banner 19, one of the car firm having advertising banner 20 and an offer from the firm having advertising banner 21.

The information to which speech recognizers and/or databases a speech input and/or recognition result is to be conveyed is given by the provider of the HTML page, while he receives the information from the customer for the advertising banners.

The provider of the HTML page can transfer information that is important for the analysis of a recognition result to the speech recognizers or databases.

The memory 25 extends the arrangement in that with successive inquiries, the speech input is stored in the memory 25. It is alternatively possible to have this memory 25 store the already generated recognition result. In that case the user can successively inquire a plurality of databases, without repeating each time the speech input or also the speech recognition.

Fig. 3 shows the arrangement of a method in which a speech input is conveyed to three different speech recognizers 7, 8 and 9. The user of the objects 19, 20 and 21 is accordingly requested to utter a speech input. This speech utterance is conveyed to the speech recognizers 7, 8 and 9 for generating each a recognition result 11, 12 and 13. The speech recognizers 7-9 analyze the recognition results 11, 12 and 13 and prepare each an inquiry for the databases 14, 15 and 16. Since, on the one hand, the recognition results 11, 12 and 13 are different, because they were generated by different speech recognizers 7-9 and, on the other hand, different inquiries are generated with these different recognition results 11, 12 and 13 during the analysis, which inquiries are applied to different databases 14, 15 and 16, the user receives with the interpretation results 22, 23 and 24 returned to him on the client 2, three responses based on different databases.

When the analysis of the recognition results is carried out in the database instead of the speech recognizer, there is a further embodiment. The databases 14-16 can then make the analyses of the individual recognition results 11, 12 and 13 with key words which are specifically contained in their respective database.

In television programs, respective features with the different stations are indicated differently. For example, with one station the feature of "children's movies" could be referenced "trick movies" with another station. If a user now says that he wishes to see a trick movie, this speech input is recognized by the assigned speech recognizer and similarly interpreted in the respective database, so that the user is ultimately offered the movies referenced trick movies or children's movies by either station.

Fig. 4 shows an arrangement in which the databases 14-16 are integrated with the speech recognizers 7-9. With smaller data files it is possible to integrate the databases 14-16 with the respective speech recognizers 7-9. Furthermore, there is represented here that a

bidirectional link is made from the respective advertising banners 19-21 to the associated interpretation results 22-24 and the associated databases 14-16. It is possible that a response to an inquiry in one of the databases 14-16 is so large that a representation of the interpretation result 22-24 on the client is not wise. In such a case, for example, only the number of found responses to a speech input are sent back to the client and displayed. When the user would like to see the interpretation result 21 of the firm having, for example, advertising banner 19, he can request it and retrieve it from the database 14. These results are then displayed on the client 2.

CLAIMS:

1. A method in which an information unit (3) that makes a speech input possible is stored on a server (1) and can be retrieved by a client (2) and the client (2) can be coupled through a communications network (6) to a plurality of speech recognizers (7-9) and a user's speech input is applied to at least one speech recognizer (7-9) for the generation of a recognition result (11-13) and the recognition result (11-13) is interpreted in a plurality of independent processes and a plurality of interpretation results (22-24) are generated which are supplied to the user.

2. A method as claimed in claim 1, characterized in that the interpretation results (22-24) are automatically returned to the client (2) or retrieved by the client.

3. A method as claimed in claim 1 or 2, characterized in that the speech input is applied to a plurality of speech recognizers (7-9) in parallel for recognition results (11-13) to be generated.

4. A method as claimed in one of the claims 1 to 3, characterized in that additional software for extracting features of the speech input is executed on the client (2) and the extracted features are applied to the assigned speech recognizer(s) (7-9).

5. A method as claimed in claim 1, characterized in that the information unit (3) is realized as an HTML page (3) and a plurality of objects (19-21) are found on one HTML page (3), which objects make a speech input possible while each object (19-21) is combined with a speech recognizer (7-9).

6. A method as claimed in claim 5, characterized in that additional information (4) for combining the objects (19-21) with a respective one of the speech recognizers (7-9) is assigned to the objects (19-21) by the server (1).

7. A method as claimed in one of more of the claims 1 to 6, characterized in that a speech input or the recognition result (11-13) is buffered in a memory (25) in order to successively execute a plurality of interpretation processes based on the buffered data.

8. A server (1) on which an information unit (3) is stored that makes a speech input possible, which information unit (3) can be retrieved by a client (2), while there is provided that

- the client (2) can be coupled to one or more speech recognizers (7-9) for generating a plurality of interpretation results (11-13) sent to a user and
- a speech input is applied to at least one speech recognizer (7-9) for generating recognition results (11-13) and for interpreting the recognition results (11-13) in a plurality of independent processes, and

for determining a combination of an object that makes a speech input possible with a speech recognizer (7-9) for generating a recognition result (11-13), additional information (4) is assigned to the object (19-21).

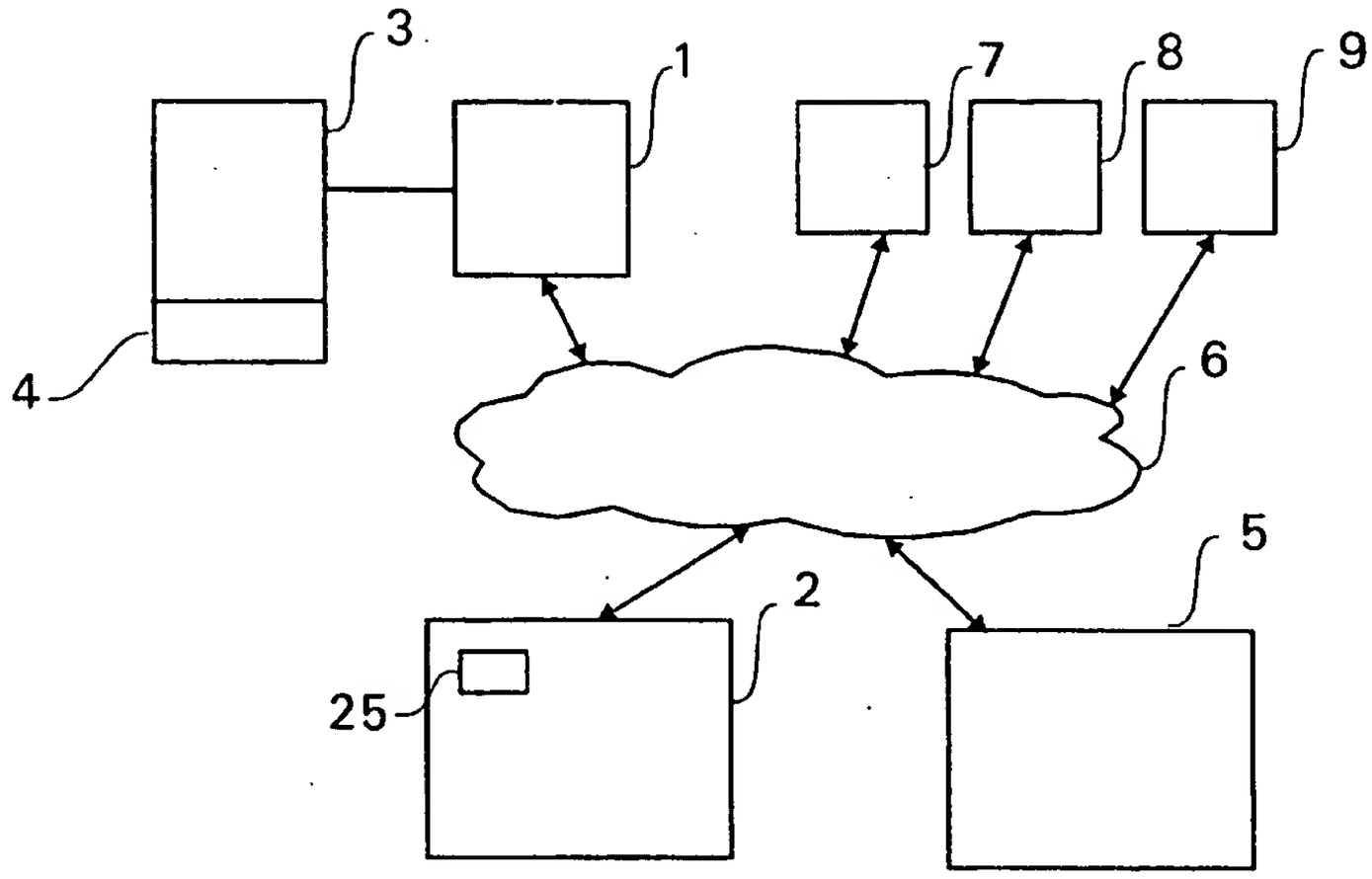


FIG. 1

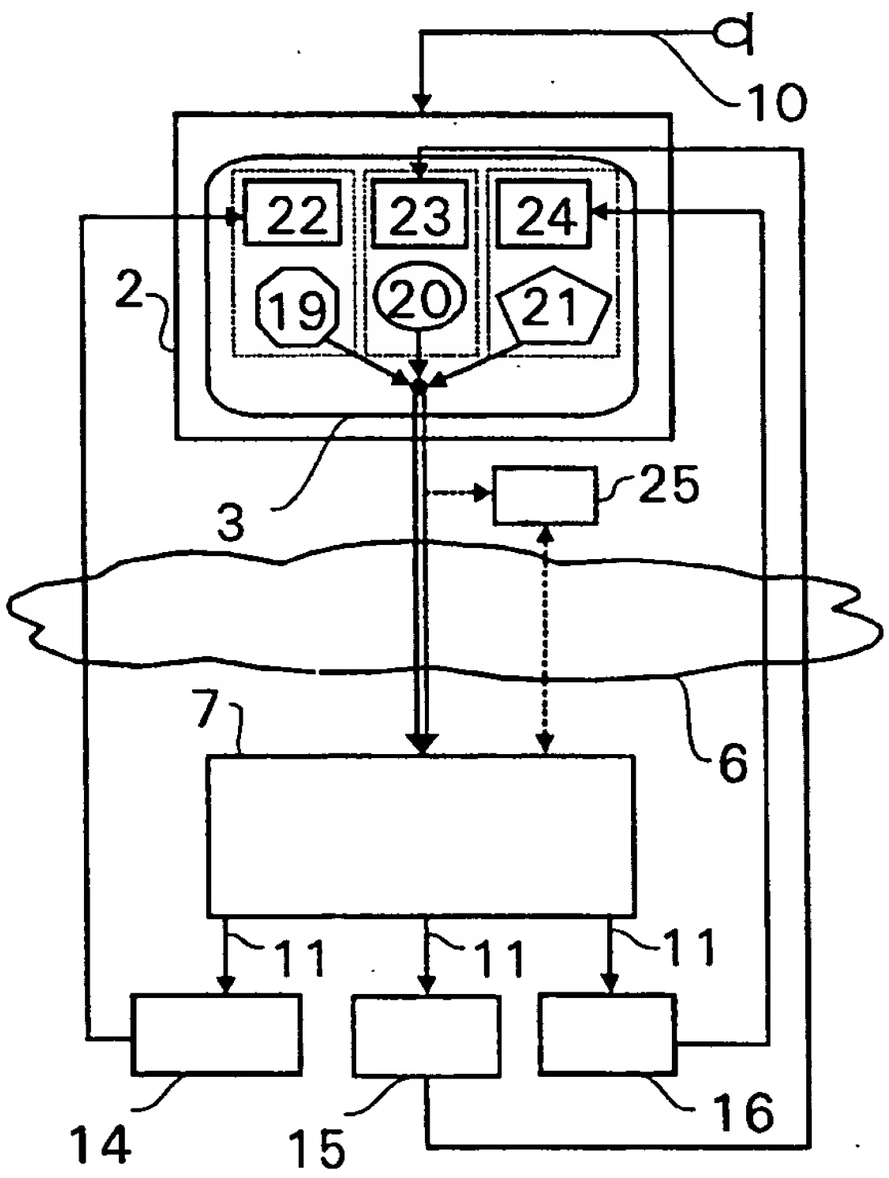


FIG. 2

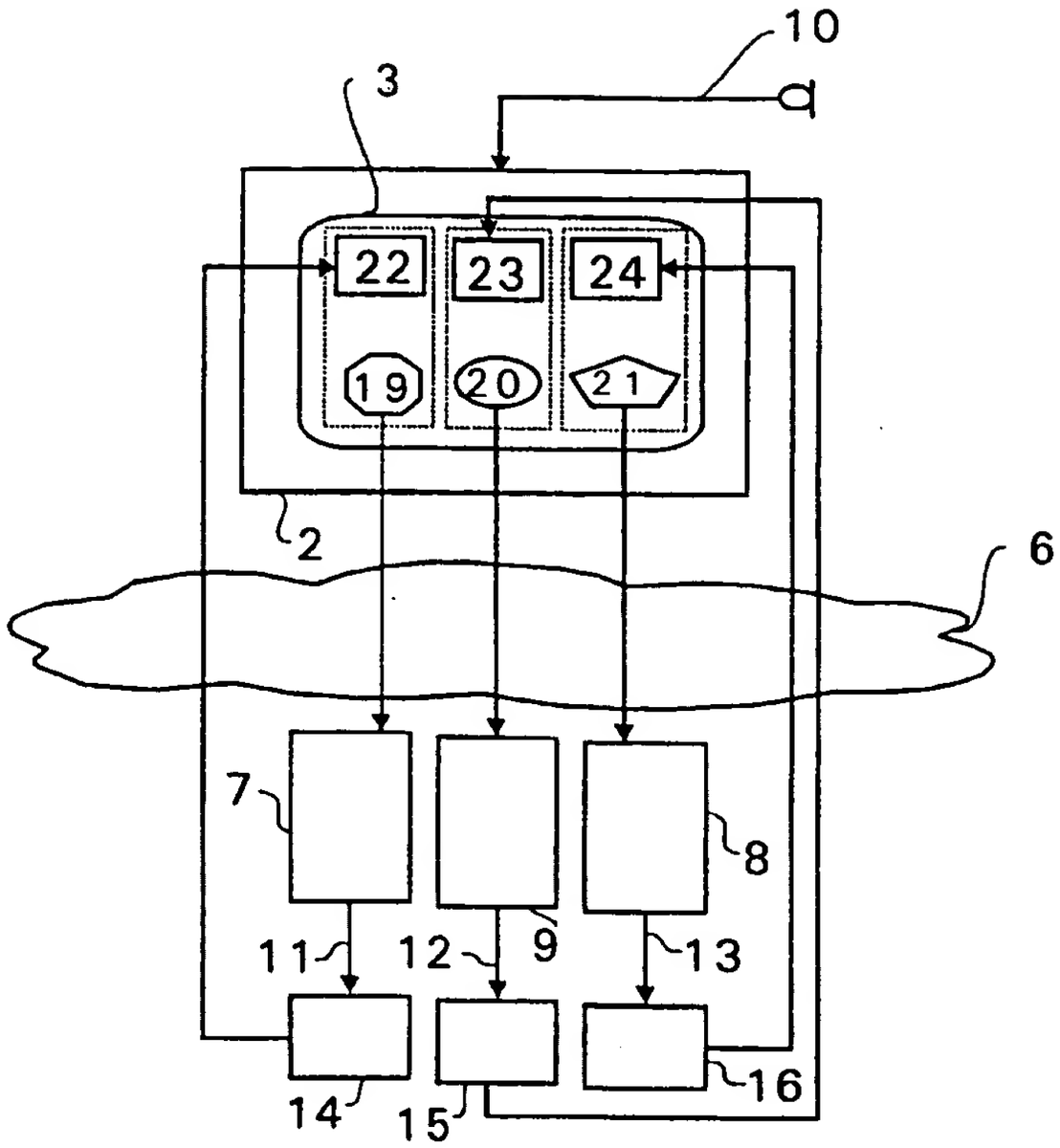

10
a
3
22
23
24
19
20
21
2
6
7
8
9
11
12
13
14
15
16

FIG. 3

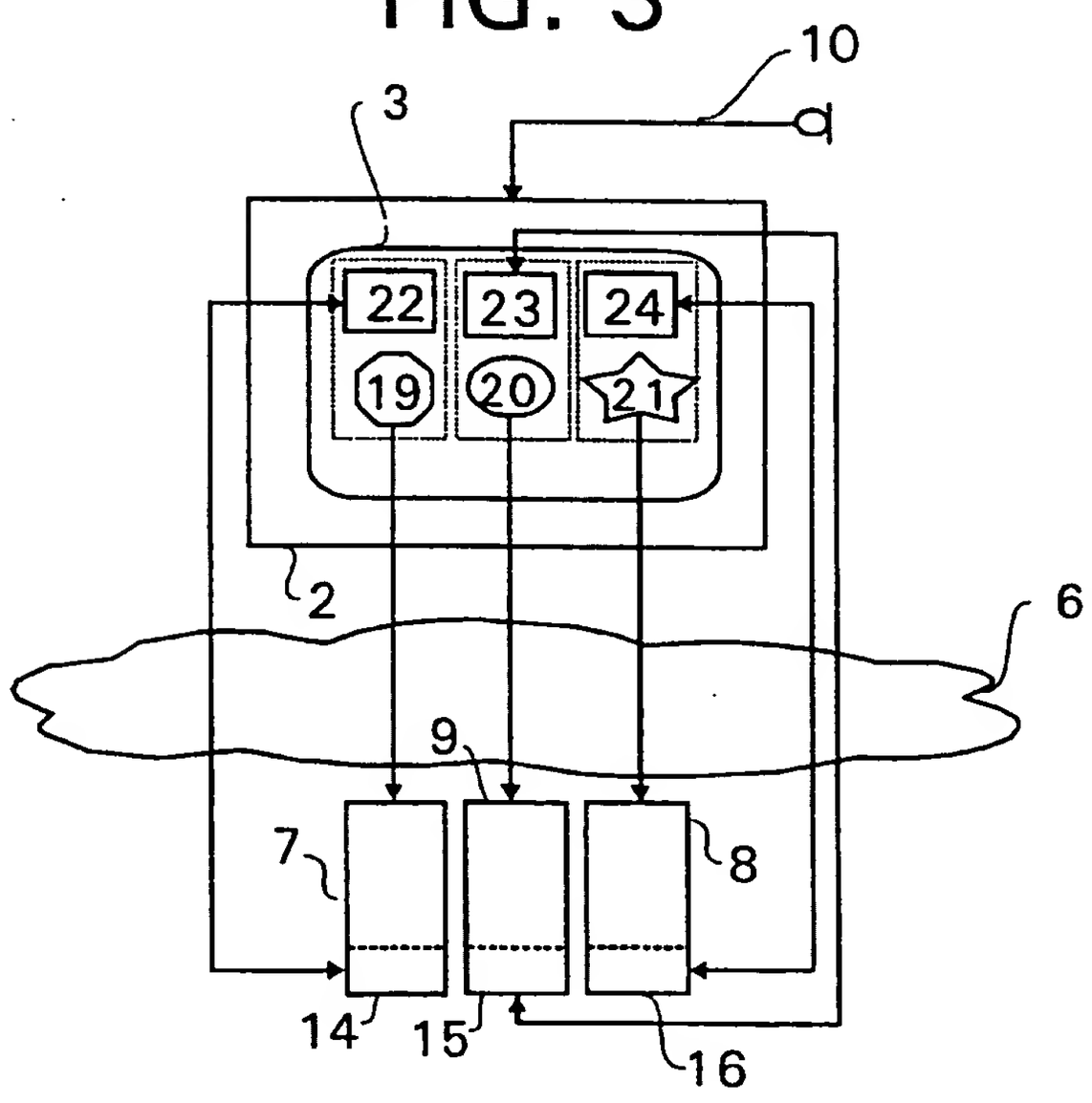

10
a
3
22
23
24
19
20
21
2
6
9
7
8
14
15
16

FIG. 4